

# Segmentation of Child-Directed Speech: A Statistical Approach

Natalya Muzinich  
Indiana University  
panteley@indiana.edu

## ABSTRACT

This paper describes how distinctive features that classify speech sounds emerge from statistical analysis of Russian child-directed speech. The analysis is based on transcriptional representation of individual speech sounds. From the analysis of bigram distribution major natural classes such as consonants and vowels further subdivided into non-palatalized versus palatalized consonants and front versus non-front vowels can be computed. The results in the form of a probabilistic FSA exhibit strong associations between the uncovered subclasses of consonants and vowels that are supported by traditional linguistic analysis.

## Categories and Subject Descriptors

I.2.7. [Artificial Intelligence] Natural Language Processing. Language parsing. I.5.3. [Pattern Recognition] Clustering. I.5.4. [Pattern Recognition] Applications. Text processing.

## General Terms

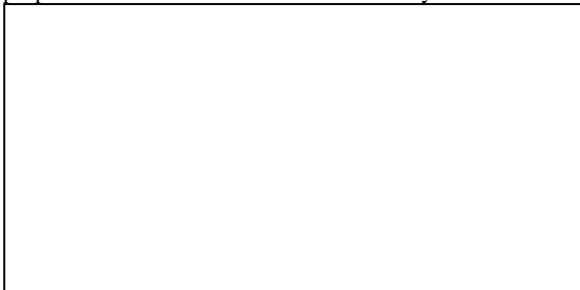
Algorithms, design, measurement, theory.

## Keywords

Principal component analysis. Singular value decomposition. Ward's method.

## 1. INTRODUCTION

How very young infants learn to segment a continuous stream of speech into meaningful units such as phrases, words and morphemes, is an important problem in language acquisition that currently lacks a definitive account. Because infants as young as several months of age are sensitive to statistical properties of sound pattern distribution in the input [11], this sensitivity has been interpreted as evidence for their use in guiding language acquisition. In addition, it has been proposed that in learning to segment a stream of child-directed speech infants rely on multiple cues to which they show sensitivity [6] The extent of reliance on each of these cues is not addressed because it is unclear how it can be measured. This paper describes a statistical approach that derives a hierarchical structure of speech sounds from the textual representation of child-directed speech based on the distributional properties of co-occurrences of individual symbols.



The resulting groupings of symbols closely correspond to the natural classes of speech sounds that are products of traditional linguistic analysis. The unifying feature within each class can be interpreted as one of the cues such as stress to which the infants show sensitivity.

Although textual representation of speech is an idealization which transforms a raw acoustic signal into discrete symbols, the psychological reality of perceiving speech as a linear sequence of discrete units is rarely disputed. Furthermore, the approach taken here does not imply that individual speech sounds are the units of infants' perception. The results exhibit strong coupling between different sound classes supporting grouping of individual sounds into syllables.

## 2. DATA

Textual representation of child-directed speech analyzed in present work comes from the CHILDES corpus. The author transcribed samples of Russian child-directed speech using Latin alphabet and adapting standard transcriptional symbols to maintain ASCII single character representation for individual sounds. 0 and 1 mark the beginning and the end, respectively, of an adult's turn in their conversation with the child, while all other word boundaries were removed. The input data file consists of 40749 characters. 49 distinct transcriptional symbols were employed.

## 3. METHOD

A 49-by-49 matrix with the rows representing the preceding speech sounds and the columns - the following speech sounds was constructed. The initial data in the matrix was raw bigram count in the input file. The standard technique of log+1 transform was performed to eliminate zero counts and scale down individual sound frequencies. By-column normalization converted the logarithms of counts to their Z-scores:  $Z = (H_c - \text{Mean}H_c) / \text{Stddev}H_c$ , where for a given following speech sound H,  $H_c$  designates its count in the input data,  $\text{Mean}H_c$  and  $\text{Stddev}H_c$  are the mean and the standard deviation, respectively, of the count of this sound in the data. High positive Z-scores in a column representing the sound  $a$  indicate the sounds that precede  $a$  unusually frequently. Low negative Z-scores designate the sounds that markedly rarely follow  $a$ . Similarity factors were derived from the normalized matrix via singular value decomposition, or SVD, described in the next section.

### 3.1. Vector Space and Latent Structure Modeling

Vector representation suggests high-dimensional space for modeling of the concept of similarity as either distance, such as Euclidean, or direction such as cosine. The current project adheres to the view under which the count correlations signify mutual dependency of both preceding and following sounds on a set of orthogonal factors. These latent variables constitute a coordinate system in a single high-dimensional space for both the preceding and the following sounds. Each dimension makes a largest possible new dissection in the cloud of data points capturing the maximum of

the variance in the remaining unaccounted for so far data. If the data are structured, a small subset of earlier dimensions explains most of the variance in the data, while the remaining ones contribute very little to the data distribution. The method, known as the principal component analysis, or PCA, performs a significant dimensionality reduction which retains those factors that impose structure onto the data and discards others that account for the noise and distortions. A noticeable change in the data variance of two adjacent factors is one way of deciding how many principal components include in explaining the uncovered structure.

This project uses singular value decomposition (or SVD), to compute the principal components. SVD is a linear algebra technique that decomposes a matrix  $X$  into a product of three matrices  $X=USV^T$ .  $S$  contains singular values measuring variance along each dimension;  $U$  contains left singular vectors, which are the coordinates in the high-dimensional space for the preceding sounds,  $V^T$  – right singular vectors, or coordinates, for the following sounds. The standard deviation, or the error covered by the principal components, is re-introduced into the data to compute factor score matrices  $U_e = \sqrt{S} * U$  and  $V_e = \sqrt{S} * V^T$ . Principal components not only minimize total entropy, that is, they maximize information content of the latent variables. They also maximize Euclidean distance and minimize the mean squared error criterion [3], which explains the choice of the similarity metric and the clustering method (Ward's) in the present study. Modeling similarity as Euclidean distance between the vectors, the project clustered them with Ward's method that minimizes the variance within each individual cluster by minimizing the sum of squares error. Ward's algorithm minimizes variance at every step of analysis and does not backtrack to change an earlier assignment to clusters to achieve greater error minimization at a later step. Branch length reflects variance as the values of the sum of squares error function  $E$ . It grows nonlinearly as the number of clusters becomes smaller, which a classification interpretation needs to take into account. Hierarchical clustering analysis was performed on both factor matrices to classify both the preceding and the following sounds.

### 3.2. Implementation

The transcription of Russian data and input file generation was carried out on a Macintosh power PC. Principal component analysis and clustering was performed on a Windows PC with 1 GB of RAM and 3 GHz Intel Pentium 4 processor.

## 4. RESULTS

The scree plot of the SVD results reveals 4 principal components accounting for about 80% of variance in the data.

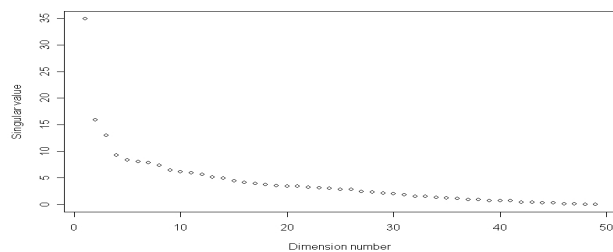


Figure 1 Scree plot of eigenvalues

From the matrix  $U_e$ , that classifies the preceding sounds dimension 1 separates consonants from vowels, while dimension 2 subdivides consonants into non-palatalized versus palatalized (i.e. with the secondary co-articulation of raising the tongue body to the roof of the mouth, as in the initial sound of the word 'yes'), a major distinction in the Russian sound system.

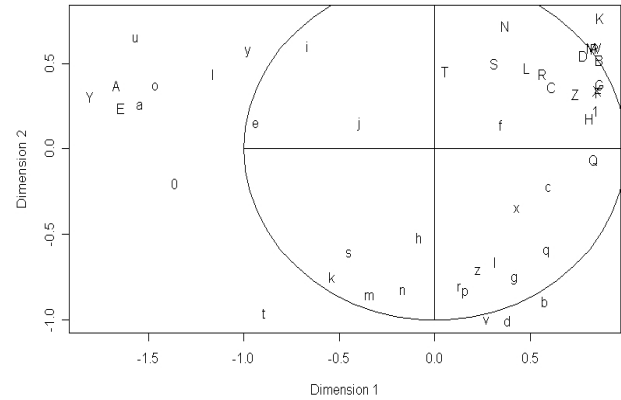


Figure 2 Major sound classes

The beginning of the separation of the front from the non-front vowels can be observed on dimension 3 while 4 isolates utterance boundaries 0 and 1. This plot is based on the  $V_e$  matrix that characterizes the following sounds on the basis of the distribution of the preceding ones. Because vowels are likely to follow consonants, this matrix represents them more accurately in the high-dimensional space.

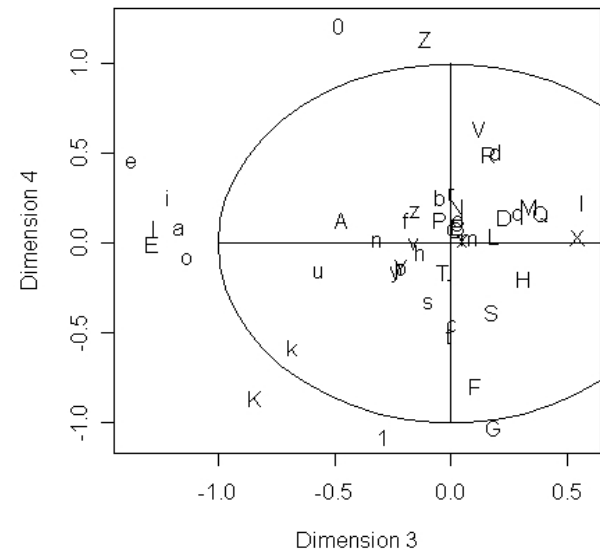


Figure 3 Beginning of vowel separation and utterance borders

Ward's clustering method of the first four principal components demonstrated the classification of consonants into palatalized and non-palatalized ones from the  $U_e$  matrix and vowels into front and non-front ones from the  $V_e$  matrix.

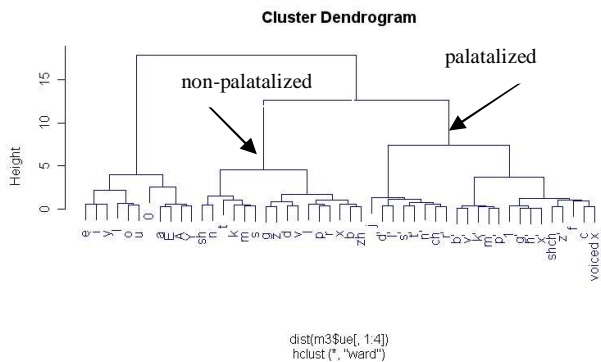


Figure 4 Consonant classification

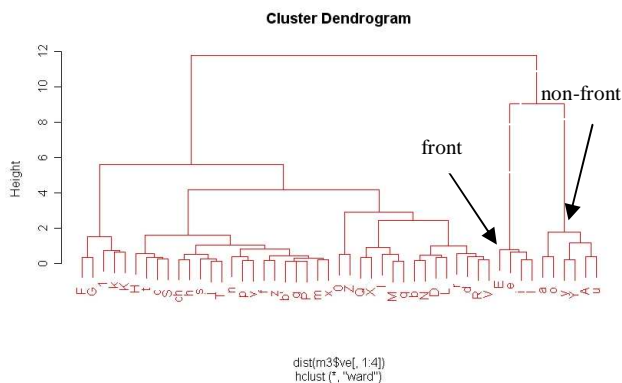


Figure 5 Vowel classification

Factor analysis and clustering produced five major categories of symbols: regular consonants (c), palatalized consonants (pc), front vowels (fv), non-front vowels (nfv) and edges. The original data file was re-labeled with the derived categories and the 5X5 matrix was constructed with category bigram counts. The size of this matrix allows for a simpler analysis of inter-category association such as  $\chi$ -square test [9], the residuals from which are presented in the table below.

Table 1 Sound categories  $\chi$ -square analysis' residuals

	c	edge	fv	nfv	pc
c	-30.33944	-4.304506	-37.922231	83.536359	-27.353085
edge	17.48195	-10.933174	-6.048521	-5.281571	-6.297925
fv	28.54727	5.549226	-23.343014	-34.820343	20.626652
nfv	33.36521	11.088931	-31.254386	-49.854695	36.257387
pc	-33.41559	-6.498834	113.611496	-20.447196	-21.333746

A high positive residual indicates that its row category is likely to be followed by its column category. The strongest association is between the palatalized consonants that are unusually highly likely to be followed by the front vowels. The following FSA is a graphic representation of the likely inter-category associations.

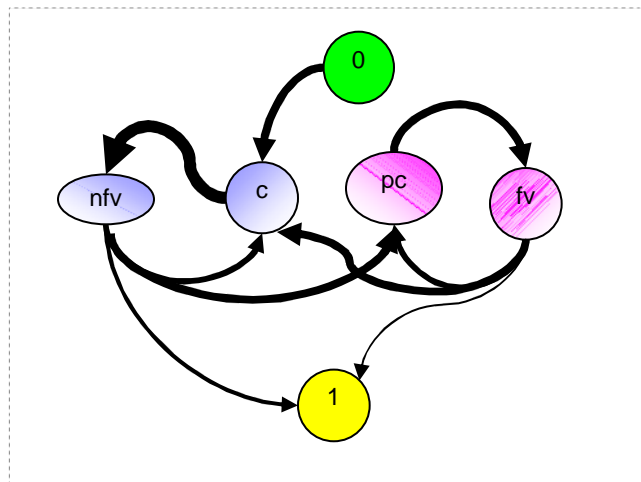


Figure 6 Associations between categories

The contextual window of two symbols most strongly captures the associations between the regular consonants which are most likely to be followed by the non-front vowels and the palatalized consonants that are most likely to be followed by the front vowels. The non-palatalized consonants are the most likely class to occur at the beginning and the vowels are the most likely class to occur at the end. Although Russian has complex syllabic structure with consonant clusters being common, the bigram context provides no support for the inner-consonantal link. The results where CV is the most statistically significant association can be used to support theories where syllables are the initial units of speech perception and to explain the prevalence of CV productions in very young children.

This approach is not at odds with more traditional text segmentation methods that use Markov models and entropy maximization [5, 7, 12]. More importantly, it does not model language acquisition or cognition directly since it is not grounded in psychological reality or experimentation. It only aims at identifying characteristics of linguistic input that are strongly associated with boundary placement. Estimated parameters can be used to predict boundaries in the new input and the model becomes a parser in automated online text segmentation and unsupervised discovery of natural sound classes.

## 5. REFERENCES

- [1] Agresti, Alan. *Categorical Data Analysis*. 2nd edition. Wiley. 2002
- [2] Agresti, Alan. *Analysis of Ordinal Categorical Data*. Wiley. 1984
- [3] Basilevsky, A. *Statistical factor analysis and related methods: theory and applications*. New York: J. Wiley. 1994.
- [4] Crawley, Michael. *Statistics. An Introduction Using R*. Wiley. 2005.
- [5] Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Lan-*

*guage Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall. 2000

- [6] Jusczyk, Peter. *The Discovery of Spoken Language*. A Bradford Book, MIT Press, Cambridge, Massachusetts, London, England. 1997.
- [7] Mccallum, Andrew, Dayne Freitag, Fernando Pereira. *Maximum Entropy Markov Models for Information Extraction and Segmentation*. Proc. 17th International Conf. on Machine Learning. 2000.
- [8] Manning, Chris and Hinrich Schütze.. *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA. 1999
- [9] Paolillo, John *Analyzing Linguistic Variation: Statistical Models and Methods*. CSLI Publications. 2002
- [10] Doug Beeferman, Adam Berger, John Lafferty. *Statistical Models for Text Segmentation*. Machine Learning 34(1-3), 117-210. 1999.
- [11] Saffran, Aslin and Newport. *Statistical cues in language acquisition: Word segmentation by infants*. COGSCI-96, 376-380, 1996.
- [12] Ratnaparnakhi.. *A Linear Observed Time Statistical Parser Based on Maximum Entropy Models*. Proceedings of the Second Conference on Empirical Methods in Natural Language Processing. UPenn, 133-142, ACL. 1997.